International Journal of
Advances in Scientific Research and Engineering (ijasre)

E-ISSN : 2454-8006

DOI: http://doi.org/10.31695/IJASRE.2018.32756

Volume 4, Issue 7
July - 2018

# A Hierarchical Framework for the Management of Virtualized Resources for Multi-Tier Services and Applications in Cloud Computing

I.M Adamu[1], Dr A.Y Gital[2] and Prof S. Boukari[3]

[1]Department of Computer Science, Federal Polytechnic Bauchi, Nigeria

[2-3]Department of Mathematical Science, Abubakar Tafawa Balewa University Bauchi

Nigeria

## ABSTRACT

*Cloud computing has become a new age technology that has got huge potentials in enterprises and markets. It makes it possible to access applications and associated data from anywhere. Organizations are able to rent resources from the cloud for storage and other computational purposes to reduce cost (operational cost). Further, they can make use of wide access to applications, based on the pay-as-you-go model. Hence there is no need for getting licenses for individual products. However one of the major pitfalls in cloud computing is related to optimization of cloud virtualized resources. Because of the uniqueness of the cloud model, resource allocation is performed with the objective of optimizing computing, communication and storage resources. Many researchers propose different methods of optimizing cloud resources, some of which are the use of soft computing approaches. Many of these optimization algorithms proposed are still under investigation and modification to improve their capabilities in managing well, the computing, communication and storage resources in the cloud. For efficient virtualized resource management in the cloud, the existing (soft computing approaches) resource allocation algorithms are becoming inefficient with the increase in the size of cloud data. In view of the above challenge,this research therefore proposes Hierarchical Framework of models for the adaptive management of virtualized resources for efficient resource allocation in cloud computing.*

**Keywords:** Cloud Computing, Virtual Machine, Cloud Services, Resource allocation Strategies, Service Level Objective.

## 1. INTRODUCTION

In recent years, the popularity, rapid growth in processing, storage technologies and success of the Internet, contributed to computing resources becoming more cheaper, powerful, and ubiquitously available than ever before [1,2]. This technological trend is popularly known as cloud computing and has led to an evolutive way to provide a solution to current and future information and communication technology (ICT) requirements [3, 4]. It is one of the upcoming computing paradigm where applications and data services are provided over the Internet [5]. At this time, most of the business organizations and educational institutions use cloud environment for data storage and other computational task such as CPU and memory usage [6]. It is also service focused to provide high quality and low-cost information services by pay-per-use model in which guarantees are offered by the cloud service providers [7]. Cloud computing is recently a booming area and has been emerging as a commercial reality in the information technology domain and thus, it is the use of computing resources such as hardwares and softwares which are delivered as a service over the internet to the customers in various organization[8].

Customer's access the cloud services through internet by using Mobile devices, PC and PDA, and Service providers provide the service to the customer [9]. These services are Infrastructure as a service (IaaS), refers to the sharing of hardware resources for executing cloud requests, typically using virtualization technology [10]. Others are Platform as a Service (PaaS) approach where offering includes a software execution environment, such as an application server, and Software as a Service approach (SaaS), where complete applications are hosted on the Internet [11]. Examples of applications that can be hosted in the cloud include word processing software, mathematical computational software and database management software [12]. The use of cloud computing brings about a lot of advantages, the most basic ones being lower costs, re-provisioning of resources and remote accessibility [13].

Cloud computing lowers cost by avoiding the capital expenditure by the organization in renting the physical infrastructure from a third party provider [14]. Due to the flexible nature of cloud computing, users can quickly access more resources from cloud providers when the need to expand business arises. The remote accessibility enables users to access the cloud services from anywhere at any time. To gain the maximum degree of the cloud computing benefits, the services offered in terms of resources should be allocated optimally to the applications running in the cloud. Many researchers propose different resource allocation algorithms, these algorithms scale in terms of waiting time, context switching and throughput but the algorithms still require some improvements to meet up with the actual data processing requirement in the cloud.

In view of the above challenge,this research therefore proposes hierarchical framework for adaptive management of virtualized resources for efficient resource allocation in cloud computing.

## 2.0 RELATED WORKS

"Baomin, et al [15]" proposed a job scheduling algorithm based on Berger Model with dual fairness constraints. Authors had mainly concentrated on fairness of resource allocation and cloud consumers' satisfaction to the provided services. Based on parameters like completion time and bandwidth, cloud consumers' tasks had been classified. According to the characteristics and preferences of tasks, resources were assigned to the cloud consumers. Authors implemented their algorithm by taking client task based on priority sorting, task are classified and binded to virtual machines, and finally released when they meet the fairness constraint. This algorithm was implemented on CloudSim toolkit and compared with optimal completion time algorithm. Results show that algorithm based on Berger Model is better. But Vector values of the general expectation vector is not very accurate

"Pandaba et al. [16]" proposed an algorithm by modifying the round robin resource allocation strategy.this algorithm begins with the time equals to the time of first request, which changes after the end of first request. When a new request is added into the ready queue in order to be granted, the algorithm calculates the average of sum of the times of requests found in the ready queue including the new arrival request. This needs two registers: (i)SR: To store the sum of the remaining burst time in the ready queue (ii)AR: To store avg. of the burst times by dividing the value found in the SR by the count of requests found in the ready queue. The modified algorithm performs reasonably regarding the wait time but fail in the response and turnaround time.

An analytical model for multi-tier internet services was proposed by "Urgaonkar et al. [17]" Processor sharing model is used to derive the average response time of the applications based on mean-value analysis. The solution presented does not have a closed-form because of complex modeling but the accuracy of the solution is demonstrated.

"GuiyiWei, et al. [18]", proposed game-theoretic method for fair resource allocation in cloud computing. Authors used Game Theory for QoS constrained resource allocation problem. Firstly, authors considered optimization problem for cloud services for which Binary Integer Programming method was proposed for initial optimization. Based on the initial result, an evolutionary mechanism was designed to achieve the final optimal and fair solution. In summary, authors focused on the sophisticated parallel computing problem on unrelated machines connected across the Internet. But the problem of these works is that applications are not clearly defined.

A system which can automatically scale its infrastructure resources is designed by "Ruth et al. [19]". The system composed of a virtual network of virtual machines capable of live migration across multi- domain physical infrastructure. By using dynamic availability of infrastructure resources and dynamic application demand, a virtual computation environment is able to automatically relocate itself across the infrastructure and scale its resources. This starts with a description of the live migration framework of multiple virtual machines with resource reservation technology. After which a series of experiments to investigate the impacts of different resource reservation methods on the performance of live migration in both source machine and target machine was conducted. An analysis on the efficiency of parallel migration strategy and workload-aware migration strategy was carried out. The metrics such as downtime, total migration time, and workload performance overheads are measured. But the above work considers only the non- preemptable scheduling policy.

"Khatri, [20]" proposes an improve Dynamic Round Robin model in which the median of the set of processes in the ready queue is considered as the optimal time quantum and if the median is less than 25 then its value must be considered as 25 to avoid the overhead of context switch. The first process in the ready queue is allocated to the CPU for a time interval of up to 1 time quantum. If the remaining burst time of the currently running process is less than or equal to 1 time quantum, the processor again allocated to the same process. Else the process is preempted and place at the rear of the ready queue otherwise the process will be halted. Experimental result shows that the proposed algorithm IDRR, when compared with various variants of Round Robin algorithm it produces minimal

"Kejiang Ye [21]", proposed resource reservation based live migration framework of multiple virtual machines. The target machine in the framework holds four virtual machines: Migration Decision Maker, Migration Controller, Resource Reservation Controller and Resource Monitor. Authors focused on improving the migration efficiency through live migration of virtual machines and



DOI: 10.31695/IJASRE.2018.32756

proposed three optimization methods: optimization in the source machine, parallel migration of multiple virtual machines and workload-aware migration strategy. To improve the migration efficiency authors had considered parameters like downtime, total migration time and workload performance overheads. Authors claimed that resource reservation strategy is required at source machine and target machine. The problem of the proposed method is that only load on virtual machine for migration is considered.

"Amit et al. [22]", proposed an algorithm in a scheduler named Haizea for resource allocation policies like immediate, best effort, advanced reservation and deadline sensitive. Haizea is a resource lease manager that uses resource leases as resource allocation abstraction and implements these leases by allocating Virtual Machines (VMs). Authors main goal was to minimize resource rejection rate and reshuffle cost in order to provide all the above mentioned resource allocation policies for IaaS cloud. Authors also used two concepts named swapping and backfilling for deadline sensitive resource allocation policy. Authors mainly considered four lease parameters for their experiments: start time, duration, deadline and number of nodes. But Backfilling algorithm is not implemented yet.

"Gopalkrishnan et al. [23]" presented a model, named as Ruled Based Resource Allocation (RBRAM) which deals with the efficient resource utilization in M-P-S (Memory-Processor-Storage) Matrix Model. Authors say that resource allocation rate should be greater than resource request rate. Major components of the system are: cloud priority manager, cloud resource allocation, virtualization system manager and end result collection. To analyse the performance of the cloud system authors considered the Cloud Efficiency Factor. However, authors also identified other parameters of Cloud System for future work.

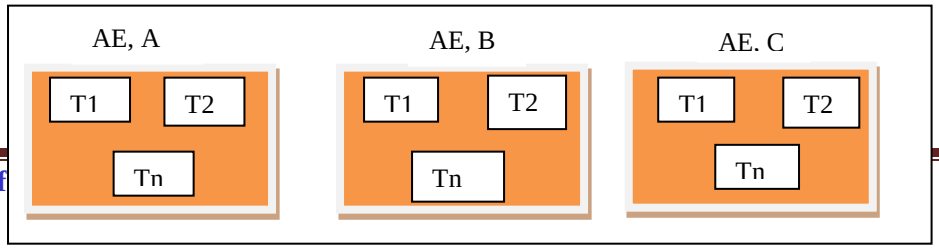
There are many proposals that dynamically manage VMs in IaaS by optimizing some objective function such as minimizing cost function, cost performance function and meeting QoS objectives. The objective function is defined as Utility property which is selected based on measures of response time, number of QoS, targets met and profit etc [24]. For multitier cloud computing systems (heterogeneous servers), resource allocation based on response time as a measure of utility function by considering CPU, memory and communication resources is proposed by "*HadiGoudarzi et al.* [24]" in which Servers are characterized based on their capacity of processing powers, memory usage and communication bandwidth. For each tier, requests of the application are distributed among some of the available servers. Each available server is assigned to exactly one of these applications tiers i.e. server can only serve the requests on that specified server. Each client request is dispatched to the server using queuing theory and this system meets the requirement of SLA such as response time and utility function based on its response time. It follows the heuristics called force-directed resource management for resource consolidation. But this system is acceptable only as long as the client behaviors remain stationary.

Resource Allocation strategies are proposed based on the nature of the applications. In the work by "Tram et al. [25]", Virtual infrastructure allocation strategies are designed for workflow based applications where resources are allocated based on the workflow representation of the application. For work flow based applications, the application logic can be interpreted and exploited to produce an execution schedule estimate. This helps the user to estimate the exact amount of resources that will be consumed for each run of the application. Four strategies such as Naive, FIFO, Optimized and services group optimization are designed to allocate resources and schedule computing tasks. Real time application which collects and analyzes real time data from external service or applications has a deadline for completing the task. This kind of application has a light weight web interface and resource intensive back end. To enable dynamic allocation of cloud resources for back-end mashups, a prototype system is implemented and evaluated for both static and adaptive allocation with a test bed cloud to allocate resources to the application. The system also accommodates new requests despite a-priori undefined resource utilization requirements. This prototype works by monitoring the CPU usage of each virtual machine and adaptively invoking additional virtual machines as required by the system.

## 3.0 METHODOLOGY

The proposed research will focus on improving the work of "Pandaba et al. [16]", by injecting mechanisms that will monitor and control the bandwidth, load and the network I/O performance when a workload is distributed to applications using the proposed modified round robin algorithm in a multi-tier environment. We aim to archive this by using an algorithm based on force-directed search to balance the load on multiple servers in order to solve the problems of high response time. And also an Adaptive Manager that dynamically adjusts multiple virtualized resource utilization to achieve application Service Level Objective (SLO) using feedback control theory.

### 3.1 Proposed Model Diagram




Multi-tier Application Environment

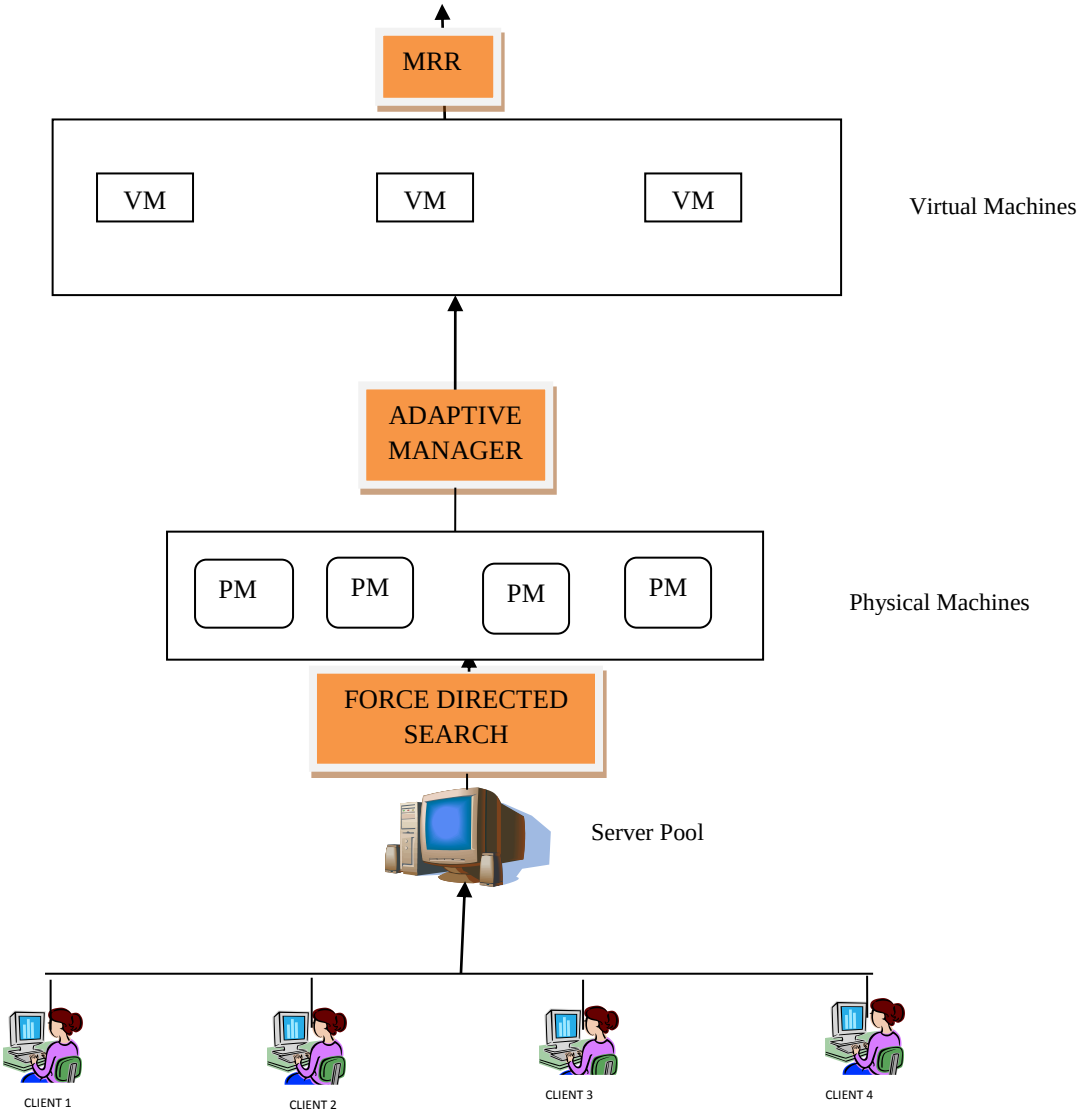



**Figure 1: proposed model diagram**

From figure 1 above, we can say that the proposed frame work consists of four major models. These models are modified round robin algorithm, which will be used to examine how the system behaves by injecting a synthetic workload measured in terms of requests per second which is distributed to applications. An adaptive manager which is propose to be a multi-input, multi-output (MIMO) resource controller with the ultimate goal to regulate multiple virtualized resources utilization to achieve application SLOs using the control inputs per-VM CPU, memory and I/O allocation of dynamic state-space feedback control system. A multi-tier model which is propose to handle applications with an arbitrary number of tiers by modeling of request processing at individual tiers and the flow of request across the tiers. and A force-directed search algorithm which is propose to consolidate resources, and further optimize the resource assignment to balance the load on multiple servers in order to solve the problems of high response time.

### 3.1.1 Modified Round robin Algorithm

This algorithm will consider the quantum to be equal to the burst time of first request, which dynamically changes after the request is executed. When a new request is added into the ready queue in order to be granted, the algorithm calculates the average of sum of the times of all the requests found in the ready queue including the request that arrived newly, i.e. quantum=totalCPUTime/noOfJobs.Where totalCPUTime is the sum of remaining CPUTime after each iteration, and noOfJobs represents the total no. Of active jobs (i.e. jobs having CPUTime greater than 0). two registers will be needed: (i)BR: To store the sum of the remaining burst time in the ready queue (ii)ABR: To store the average value of the burst times which will be calculated by dividing the value found in the BR by the total number of requests found in the ready queue. After execution, if a request finishes its burst time, then it will be removed from ready queue. BR will be updated by subtracting the time consumed by this



DOI: 10.31695/IJASRE.2018.32756

request. Also, if the remaining CPU burst time of the currently running process is less than time quantum then the CPU will be allocated again to the currently running process for remaining CPU burst time. After completion of execution, remove the process from the ready queue and allocate CPU to next process. ABR will be updated according to the burst time of the new request. While the readyqueue is not empty, average waiting time and average turnaround time will be calculated.

### 3.1.2 Adaptive Manager

In our management system model, the datacenter would consists of a set of physical machines (PM) each hosting multiple VMs through a hypervisor. We will assume that the number of physical machines is fixed and that they all belong to the same cluster with the possibility to perform a live migration of a VM between two arbitrary PMs. An *Application Environment* (AE) contains an application hosted by the cloud system. An AE is associated with specific performance goals specified in a SLA contract. An AE can embed an arbitrary application topology which can span over one or multiple VMs (e.g. multi-tier Web application, master-worker grid application). We will also assume that a running VM is associated with one and only one AE. Applications cannot request a VM with an arbitrary resource capacity in terms of CPU power and memory size. The VMs available to the application would be chosen among a set of pre-defined VM classes. Each VM class will have a specific CPU and memory capacity e.g. 4GHz of CPU capacity and 2GB of memory. Also, each AE would be associated with an application-specific Black Box Unit (BBU), which will evaluate the chances of allocating more VMs or releasing existing VMs to/from the AE on the basis of the current distributed workload by the proposed modified round robin algorithm using service-level metrics (response time and number of requests per second). The main job of the black box unit is to compute a utility function which will give a measure of application satisfaction with a specific resource allocation (CPU, RAM) given its current workload and SLA goal. The black box units would interact with a *Decision control box* (DCB) which will be the decision-making module within the control loop. The DCB would be responsible for arbitrating resource requirements coming from every AE and treats each black-box unit without taking into consideration of the nature of the application or the way the BBU computes its utility function. The DCB would receive as input (i) the utility functions from every BBU and (ii) system-level performance metrics (e.g. CPU load) from virtual and physical servers. The output of the DCB would consists of management actions directed to the server hypervisor and notifications sent to BBUs, which later notifies the BBU that (i) an existing VM has been upgraded or downgraded, i.e its class and resource capacity has been changed, (ii) a new VM with a given resource capacity has been allocated to the application and (iii) a VM belonging to the application is being preempted and that the application should relinquish it promptly. Management actions would also include the life-cycle management of VM (starting, stopping VMs) and the trigger of a live migration of a running VM.

### 3.1.3 Force Directed Search Algorithm

We assume that the cloud computing system has a central manager that has information about clients and servers. Each client will be identified by a unique id, represented by index *i*. Each server will similarly be identified by a unique id, denoted by index *j*. There are often a set of application tiers that an application needs to complete. For each tier, requests of the application would be distributed among some of the available servers. Each ON server would be assigned to exactly one of these application tiers. This means that if a server is assigned to some tier, it can only serve the requests on that specified tier. Each application has a constraint on memory allocation in each tier. This means that a constant amount of memory should be allocated to the ith client in each server that serves a portion of the client's requests in tier t. No hard constraints are imposed on the processing and communication resource allocations but these allocations determine the system profit.

### 3.1.4 Multi-tier service model

In the *Multi-tier service model,* we Consider $i^{th}$ client with an ordered set of application tiers. This ordered set is a subset of the available tiers in the cloud computing system. The inter-arrival time of requests for the $i^{th}$ client is assumed to follow an exponential distribution with rate parameter. In addition, in each level of the application tier, the client's requests would be distributed among a number of servers. For each tier, there is a probability $p_i^t$ that the requests will not go to the next application tier and instead return to the previous tier. Therefore the requests would be moving in two different directions: forward and backward. The backward requests would be served by the servers that previously served those requests in the forward direction, and because the backward streams of requests may have different service times, they will be put in different queues. In this model, the requests in the backward direction go to the previous tier with probability of one.

## 4.0 CONCLUSION

A hierarchical management framework for virtualized resource allocation in cloud computing was proposed. At the lower level, a force-directed search algorithm which will consolidate resources, and further optimize the resource assignment to balance the load on multiple servers in order to solve the problems of high response time was proposed. At the next level, an adaptive manager which is a multi-input, multi-output (MIMO) resource controller with the ultimate goal to regulate multiple virtualized

resources utilization to achieve application SLOs using the control inputs per-VM CPU, memory and I/O allocation of dynamic state-space feedback control system was proposed. While at the upper level, a hybrid round robin algorithm which will be used to examine how the system behaves by injecting a synthetic workload measured in terms of requests per second which is distributed to applications was proposed. The cooperation of these control levels will ensure the achievement of QoS metrics and the satisfaction of resource constraints under highly varying workload.

## 5. EXPECTED RESULTS

The proposed system is expected to achieve the following:

1. A model for improving cloud data processing requirement in terms of response time, network I/O performance and bandwidth.
2. A model for optimal resource utilization in the cloud to avoid unnecessary resource wastage.
3. A resource controller that regulate multiple virtualized resources utilization within the cloud.
4. To show the effectiveness of the model and its weaknesses for further research.

## REFERENCE

[1] *Manyika, James,* The Internet of Things: Mapping the Value beyond the Hype, *McKinsey Glob Institute, 2015 p. 4.* http://www.mckinsey.com/insights/business_technology/the_internet_of_things_the_value_of_digitizing_the_physical_world

[2] Baguley, Richard, and Colin McDonald, Appliance Science: The Internet of Toasters (and Other Things).*http://www.cnet.com/news/appliance-science-the-internet-of-toasters-and-other-things* 2015

[3] Fernando, Niroshinie, Seng Loke W., and WennyRahayu, Mobile cloud computing: A survey, *Future Generation Computer Systems*, 29(1), 2013 pp. 84-106.

[4] Sasikala, P. Research challenges and potential green growth technological applications in cloud computing, *International Journal on Cloud Computing*, 2(1), 2013 pp. 1-19,

[5] Mohammad-Hossein Malekloo, Nadjia Kara, May El Barachi, An energy efficient and SLA compliant approach for resource allocation and consolidation in cloud computing environments, ***Elsevier*** 2017 **pg 9-24**

[6] Mohammad Sajid, ZahidRaza, Cloud Computing: Issues &Challenges. International Conference on Cloud, Big Data and Trust Nov 13-15, 2013 *RGPV*

[7] Zhang Q., Cheng L. and R. Boutaba, Cloud Computing: State-of-the-Art and Research Challenges. *Journal of Internet Services and Application*, 2010 vol. 1, no. 1.

[8] Kasture A Hardware and Software Architecture for Efficient Datacenters PH. D. Thesis in the Department of Electrical Engineering and Computer MIT 2017.

[9] Vincent C. Emeakaroha, IvonaBrandic, Michael Maurer, Ivan Breskovic, SLA-Aware Application Deployment and Resource Allocation in Clouds. *35th IEEE Annual Computer Software and Application Conference Workshops,* 2011 pp. 298-303.

[10] Krishnaveni N., Sivakumar G**.,** Survey On Dynamic Resource Allocation Strategy in Cloud Computing Environment, *International Journal of Computer Applications Technology and Research* 2013 *vol 2- issue* 6,731-737.

[11 ]Zuling Kang, Hongbing Wang, A Novel Approach to Allocate Cloud Resource with Different Performance Traits. *IEEE 10th International Conference on Services Computing*, 2013 ISSN: 978-0-7695-5026-8/13, DOI 10.1109/ SCC.2013.109.

[12] Ronak Patel, Sanjay Patel, Survey on Resource Allocation Strategies in Cloud Computing. *International Journal of Engineering Research & Technology* (IJERT) 2013 Vol. 2 Issue 2, ISSN.

[13] Nikolaos Leontious and Spyros Denazos, A hierarchical control framework of load balancing and resource allocation of cloud computing services, 2018 *Elsevier* volume 67 pages 235-251

[14] Tan Xianoying, Dynamic resource allocation in cloud download service, 2017 *Elsevier* volume 24, pages 3-59

[15] BaominXu, Chunyan Zhao, Enzhao Hu and Bin Hu, Job scheduling algorithm based on Berger model in cloud environment, *ELSEVIER - Advances in Engineering Software* 2011 419–425

[16] Pandaba Pradhan, Prafulla Ku. Behera, B N B Ray Modified Round Robin Algorithm for Resource Allocation in Cloud Computing, *International Conference on Computational Modeling and Security* (CMS) 2016 878-890



DOI: 10.31695/IJASRE.2018.32756

[17] Urgaonkar B., Pacifici G., Shenoy P., Spreitzer M., and Tantawi A. An analytical model for multi-tier Internet services and its applications, *ACM International Conference on Measurement and Modeling of Computer Systems* 2011.

[18] GuiyiWei, Athanasios V. Vasilakos, Yao Zheng and NaixueXiong, A game-theoretic method of resource allocation for cloud computing services, *in Springer, J Supercomput* 2010 54: 252-269.

[19] Ruth P., Rhee J., Xu D., Kennell R. and Goasguen S., Autonomic Adaptation of virtual computational environments in a multi-domain infrastructure, *IEEE International conference on Autonomic Computing.* 2012 pp.5-14.

[20] Khatri, J. An Improved Dynamic Round Robin CPU Scheduling Algorithm Based on Variant Time Quantum, *IOSR Journal of Computer Engineering (IOSR-JCE)* , 2016 PP 35-40 .

[21] Kejiang Ye, Xiaohong Jiang, Dawei Huang, Jianhai Chen and Bei Wang, Live Migration of Multiple Virtual Machines with Resource Reservation in Cloud Computing Environments *in IEEE 4th International Conference on Cloud Computing*, 2011 978-0-7695-4460-1.

[22] Amit Nathani, Sanjay Chaudhary and Gaurav Somani, Policy based resource allocation in IaaS Cloud ELSEVIER - Future Generation Computer Systems 2011 28,94–103.

[23] Gopalkrishnan T.R., Nair, and Vaidehi M, Efficient resource arbitation and allocation stratargies in cloud computing through virtualization, *In Proceedings of IEEE* CCIS2011, 978-1-61284-204-2/11 2011

[24] HadiGoudarzi and MassoudPedram. Maximizing Profit in Cloud Computing System via Resource Allocation, *IEEE 31st International Conference on Distributed Computing Systems Workshops 2011* pp, 1- 6.

[25] Tram Truong Huu& John Montagnat, Virtual Resource Allocations distribution on a cloud infrastructure, *IEEE 2010*. pp.612-617.